Methods (supplementary section)

Clustering algorithm

Computational methodologies to cluster proteins by the creation of annotated families do not succeed in clustering the whole current sequence databases (Liu and Rost, 2003). These approaches are often designed to produce persistent items (for example permanent identifiers for clusters), which limits their ability to deal with the immense number of new protein sequences. Other automated approaches that rely on the removal of redundant sequences are fast but do not cluster all orthologs of a protein because they only aggregate sequences with high identity (Holm and Sander, 1998).

We propose an alternative approach for clustering similar protein sequences. Clustered sequences can have low but significant levels of identity, but are required to be similar along their full lengths without long unmatched segments. We want to make sure that when two sequences are clustered they have exactly the same domain distribution to increase the chances that they are functionally equivalent (Koonin and Mushegian, 1996; Ponting et al., 2000). Most importantly, this approach can be used just to compress the current database anew with each new version of it, without any attempt to keep persistent clusters.

We use the following greedy algorithm to cluster the sequences of a protein database. We first sort the sequences by decreasing length, then pick the longest sequence as the query sequence. We start by comparing with BLAST (Altschul et al., 1997) the query sequence against the subset of sequences which either have the same length as the query or are shorter by less than N amino acids (aa). We retain all sequences with BLAST matches covering the full length of both the query and the subject sequence (or shorter by less than N/2 aa from each end of both sequences). N is 30 aa for query sequences of length 200 aa or more, 20 aa for sequences shorter than 200 aa, 10 aa for sequences shorter than 100 aa, and 6 aa for sequences shorter than 60 aa and longer than 49 aa. We do not cluster sequences shorter than 50 residues. Then, sequences are considered to match the query sequence if they are sufficiently similar in sequence as detected by a BLAST single hit (either continuous or gapped) that covers all (or almost all) of the protein. An identity threshold of 24.8% was used for BLAST hits of more than 80 aa, and a length dependent threshold of 290.15 * length ** -0.562 was used for shorter hits following (Sander and Schneider, 1991). Query and matched sequences are grouped in a cluster and removed from the database. Then the longest sequence in the database is chosen as new query and the procedure is repeated until no sequences remain.

The algorithm can cluster current versions of SwissProt (version 52, March 2007, contains 252,971 sequences) in a matter of hours. CPU time needed scales quadratically with the number of sequences. For example, clustering the 3.65M sequences of UniRef100 version 8.5 (September 2006) took 2,200 hours on a single CPU of a Sun Microsystems v60x server with dual 3.06 GHz Xeon CPUs and 2GB of RAM. This is the clustered release we used for annotation analyses and in our BLAST web server. For the historical analysis of all major UniRef100 releases, we parallelized the algorithm to produce the clusters in a reasonable time. This was done by segmenting the database by protein length, clustering the segments separately, and doing a second pass to pick clusters of sequences across segments. This gives a slightly less compressed result.

Gene annotation

For each protein in the original data set we used the uniref100.xml v8.5 file to determine the taxonomic IDs. Taxonomic paths were derived based on the NCBI taxonomy database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) (Wheeler et al., 2007). For all UniRef proteins, Gene Ontology, PDB, and PubMed annotations were extracted from the UniProt Knowledge base uniprot_trembl.dat and uniprot_sprot.dat files (v9.3). PubMed IDs referenced in more than 12 protein entries were not considered as they tend to be uninformative for the purposes of characterizing particular protein sequences (Perez et al., 2004). Information about the Pfam domains was extracted from the Pfam database (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/current_release/; version 21.0, Nov. 2006). Pfam domain information for both SwissProt, Trembl and UPI proteins was extracted from the data sets provided at ftp://ftp.ebi.ac.uk/pub/databases/interpro (Mulder et al., 2007) for both UPI and UniRef identifiers. The resulting data was imported into a MySQL database and statistics were generated from queries on that database.

Domain coverage of sequences

For each cluster the sequence with most coverage by Pfam domains was selected as representative. The total length of the 1,354,861 cluster representatives was 530,978,620 aa (average of 392 aa per sequence). A total of 994,626 matches to 8,242 Pfam domains covered 167,662,952 amino acids (average of 160 aa per hit).

The set of cluster representatives with domains masked out was scanned for transmembrane regions using TMHMM2 (Krogh et al., 2001). Positive matches totaled an aggregated sequence of 12M aa. Those regions were masked and the resulting sequences were scanned for coiled coil regions using Coils (Lupas, 1997). Positive matches (13M aa) were masked and the resulting sequences were scanned for low complexity regions using the seg program (Wootton and Federhen, 1996). A further 38M aa were masked by this procedure. Finally, regions left unmasked but with length of less than 50 amino acids were considered as likely incapable of forming structured domains (possibly being linkers between domains or small C- and N-terminal regions) and were also masked (a further 50M aa). The remaining sequence fragments (112M aa, 21% of all aggregated sequence) were interpreted as forming part of as yet undetected domains. The masked 250M aa (which make up 47% of all protein representatives aggregated sequence length) correspond to regions of low sequence conservation that could be very difficult to associate with structural domains.

A total number of domains can be estimated for a given protein database size assuming that the current distribution of domain occurrences per domain (Figure S3) will scale as new domains are discovered and new matches to these domains are recorded. Given an estimation of the total of sequences amenable to description by domains, we can compute the possible number of hits to domains that would fit, D, assuming that the average length for an extended set of domains is constant. We know the current number of hits to domains, d, that is the integral of the curve depicted in supplementary Figure S3. If the surface (number of hits) scales up to D, the length of the distribution has to be multiplied by the root of D/d. Then the total number of domains could be approximated to the current end value of the distribution (8,242) times root of D/d domains.

Accessibility of method and data

We have implemented different mechanisms to query our set of clusters as a web interface named "*Bluster*" for *BLAST the Cluster,* available at http://www.ogic.ca/projects/bluster. Clusters can be retrieved by protein sequence identifier (using SwissProt or UniProt identifiers), by a combination of taxa and annotation properties, or by sequence similarity to a query protein sequence via a BLAST server which uses the cluster leader sequences as search database. The results of BLAST queries against the cluster leaders are generated more quickly and are notably more compact than searches directly against the full database which is 2.6 times larger. FASTA files of matching leaders and clusters may be downloaded from the interface. Bluster displays the taxonomic, domain and other properties of the clusters resulting from a query.

Members and annotations of UniRef100 r8.5 sequence clusters and other supplementary material are available at http://www.ogic.ca/projects/cluster/.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389-3402.

Holm, L. and Sander, C. (1998) Removing near-neighbour redundancy from large protein sequence collections. *Bioinformatics*, **14**, 423-429.

Koonin, E.V. and Mushegian, A.R. (1996) Complete genome sequences of cellular life forms: glimpses of theoretical evolutionary genomics. *Curr Opin Genet Dev*, **6**, 757-762.

Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-580.

Liu, J. and Rost, B. (2003) Domains, motifs and clusters in the protein universe. *Curr Opin Chem Biol*, **7**, 5-11.

Lupas, A. (1997) Predicting coiled-coil regions in proteins. *Curr Opin Struct Biol*, **7**, 388-393.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., Courcelle, E., Das, U., Daugherty, L., Dibley, M., Finn, R., Fleischmann, W., Gough, J., Haft, D., Hulo, N., Hunter, S., Kahn, D., Kanapin, A., Kejariwal, A., Labarga, A., Langendijk-Genevaux, P.S., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Orengo, C., Petryszak, R., Selengut, J.D., Sigrist, C.J., Thomas, P.D., Valentin, F., Wilson, D., Wu, C.H. and Yeats, C. (2007) New developments in the InterPro database. *Nucleic Acids Res*, **35**, D224-228.

Perez, A.J., Perez-Iratxeta, C., Bork, P., Thode, G. and Andrade, M.A. (2004) Gene annotation from scientific literature using mappings between keyword systems. *Bioinformatics*, **20**, 2084-2091.

Ponting, C.P., Schultz, J., Copley, R.R., Andrade, M.A. and Bork, P. (2000) Evolution of domain families. *Adv Protein Chem*, **54**, 185-244.

Sander, C. and Schneider, R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56-68.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Geer, L.Y., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Miller, V., Pruitt, K.D., Schuler, G.D., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L. and Yaschenko, E. (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **35**, D5-12.

Wootton, J.C. and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol*, **266**, 554-571.